Wayne Eckerson: Hi my name is Wayne Eckerson, I am Director of Research at TechTarget and I am here today to talk to you about “Big Data and Its Impact on the Data Warehousing Architecture”.

Big Data and DW Architectures

Wayne Eckerson
Director of Research
Applications and Architecture Media Group
TechTarget

In the last video in this series, I talked about the perfect storm, the conversions of four major trends about transforming corporate computing. The first is big data, the fact that we are capturing larger volumes of data than ever before and different types of data, largely unstructured data from our website.

The second is real-time data; we are capturing this data in real-time, delivering it to users in near real-time, so they can take actions at the points concerned and make a difference. The third is deep analytics, they were applying statistical models and machine learning models of algorithms
through these large volumes of data to tease out the hidden and relevant patterns and anomalies in the data for business gains, and the last is that we are allowing these users to query these large volumes of data, get responses back in a second rather than minutes, hours or even days as in the past.

The Perfect Storm

Big Data
+
Real-time Updates
+
Deep Analytics
+
Fast Queries

So as a data warehousing manager, the first thing you have to ask from the architectural perspective is, is your data warehousing accurate, is the infrastructure up to the challenge of big data? Can it address this perfect storm at times. You have to ask more specifically does its scale, does it load data fast enough, does it handle unstructured data and support detailed analytics; can it offer fast queries to them? If it doesn’t feel any of those things you are probably in the market for new technology.
Is Your DW Platform Up To the Challenge?

Does it:
• Scale?
• Load data fast?
• Handle unstructured data?
• Support deep analytics?
• Offer fast queries?

Fortunately there is new technology available that is specifically addressed to meet these needs. I call this analytical platform and people call it data warehousing appliances or simply just analytical databases and these are purpose built designed specifically for query processing and analytics that provide superior query performance or case performance compared to general purpose database solution.

Analytic Platforms (“DW Appliances”)

Oracle Exadata
1010data
Aster Data (Teradata)
Calpont
Dataupia
Exasol
Greenplum (EMC)
IBM SmartAnalytics
Infobright
Kognitio
Microsoft Parallel DW
Netezza (IBM)
Paraccel
Pervasive
Sand Technology
SAP Hana
Sybase IQ (SAP)
Teradata
Vertica (HP)

An integrated, hardware-software data management system designed explicitly for query processing and analysis that provides dramatically higher price/performance and availability compared to general purpose solutions.
Now it’s a bevy of companies in the marketplace who are selling analytical platforms, there has been a lot of consolidation among these vendors in the last 12 months and this testifies to the power of these platforms to deliver outsized performance gains for your data warehousing module. In short this is gain changing technology primarily because these platforms are purpose built to work analytics, they quickly deploy, they come pre-configured, no indexes, no aggregates to the design, they also have built in analytics fiber to accelerate analytics process, often faster and more scalable as you add notes performance scales and more extensible and available and most importantly from the CFO’s perspective, but less costly, that’s what power and less calling and less space requirements and fewer people required to maintain these environments.

**Game-changing technology**

- **Purpose built**
  - For analytics in general
  - For specific analytic workloads
- **Quicker to deploy**
  - Preconfigured and tuned
  - Built-in analytic libraries
- **Faster and more scalable**
  - Faster query and load times
  - Linear performance (data scalability)
- **More extensible and available**
  - Seamlessly add nodes
  - Redistribute data online
- **Less costly**
  - Less power, cooling, space
  - Fewer people to maintain

Now just when we thought we got our arms around the Big Data issues with analytical platforms, a new technology has come to fore and that’s called the doop and the doop brings new perspective and a lot of new innovative approaches to addressing our big data issues. A hadoop is an open source distributed file system and offers two major benefits for addressing big data issues. One because it is open source it shows, you don’t have to spend tens of millions of dollars to address big data issues. Second doop because it’s a distributed file system it’s ideally suited to handle large volumes of unstructured data which is the fastest growing segment.
Big data source, now this chart compares use of a doop to use traditional relational databases in analytics platforms. We can see at the top the two sets of bars at the top to show that both are being used to analyze varying levels of detailed data, but there is a divergence in the last three sets of bars. The second set shows that they are using the doop to do and perform different types of analytics largely analytics that can't be easily performed with SQL and the last two sets of bar show that a doop is being used to store and process large volumes of unstructured data. So in essence, it being used as big staging there where all this unstructured data that you really could not store and maintain cost effectively in the past.
So given this, what's the impact of the doop and analytical platforms on your data warehousing aspects, so this slide shows what I think is an emerging architecture for data warehousing. Now the bottom of the slide shows a classic traditional data warehousing architecture where we are pulling data from our operational systems, largely transactional data using ETL tools to populate the warehouse, and we are using ETL tools to populate these data marts subjects specific data marts for specific departments that casual users can access via BI tool that’s there.

So what’s new or rather what’s emerging is anything above that. So, on the left hand side we see a lot of unstructured data, documents, text, web blogs that we are pulling as files as is essentially into a hadoop cluster and we are storing that as long as we want and as long as we need. We are using a doop integrated ETL tool and to pull this unstructured data, essentially they are still aggregated and push it into the warehouse where we can then analyze and pulled on it using our traditional BI tools.
Now what's really new here is, the options that are we giving to the power user.

The power user now, if they have sufficient skills and programming capabilities can actually analyze all the raw, unstructured data and find out the doop prospects.
It can also, if the IT department has set it up, access a virtual SAN box inside the data warehouse where they can ideally load their own data or mingle it with the data warehousing data, and perform their analysis.

### Key Characteristics

- **No longer required**
  - Indexes, aggregates, partitioning, DBAs
- **Technology differentiators**
  - Proper balancing of CPU-memory-disk
  - Storage-level filtering
  - Scale out on commodity servers
  - Auto redistribution of data across nodes
  - Memory-based caches or databases
  - Columnar compression
  - Judicious use of Flash drives

Now if that type of query processing is bogging down performance for everyone else that uses the warehouse, the IT department can offload a replica of the data warehouse or a subset to an analytical platform let’s say a SAN box or specific analytical data model that power users can also
access to their heart’s content that without bogging down performance of the main data warehouse.

So in a nutshell this is the type of architecture, I see many companies are adapting to deal with their big data issues. This is Wayne Eckerson, hope you enjoyed this video, have a great day.

Architectural Use Cases

Enterprise data warehouse 68%
Dependent data mart 42%
Data staging area 42%
Independent data mart 39%

From Merv Adrian and Colin White, “Analytic Platforms: Beyond the Traditional Data Warehouse” (BeyeResearch, 2010). 223 respondents